Supervised Learning

Filter Methods

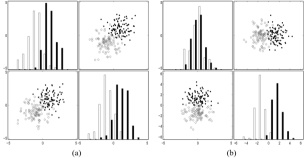
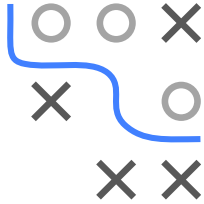


Figure 1: **Information gain from presumably redundant variables.** (a) A two class problem with independently and identically distributed (i.i.d.) variables. Each class has a Gaussian distribution with no covariance. (b) The same example after a 45 degree rotation showing that a combination of the two variables yields a separation improvement by a factor $\sqrt{2}$. I.i.d. variables are not truly redundant.

Learning goals

- Understand how filter methods work and how to apply them for feature selection.
- Understand advantages of filter methods and potential problems.
- Know filter methods based on correlation, test statistics, and mutual information.

INTRODUCTION



- **Filter methods** construct a measure that quantifies the dependency between all features and the target variable.
- They yield a numerical score for each feature x_j, according to which we rank the features.
- They are model-agnostic and can be applied generically.
- Filter methods are strongly related to methods for determining variable importance.

FILTER METHODS CAN BE MISLEADING

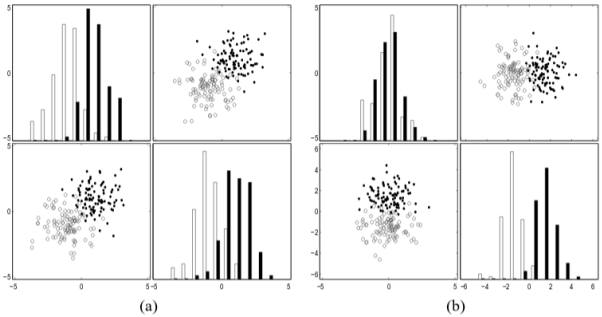


Figure 1: **Information gain from presumably redundant variables.** (a) A two class problem with independently and identically distributed (i.i.d.) variables. Each class has a Gaussian distribution with no covariance. (b) The same example after a 45 degree rotation showing that a combination of the two variables yields a separation improvement by a factor $\sqrt{2}$. I.i.d. variables are not truly redundant.

Isabelle Guyon, André Elisseeff (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (3) p. 1157-1182.

FILTER METHODS CAN BE MISLEADING

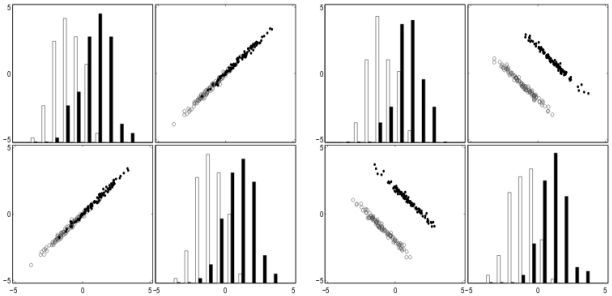
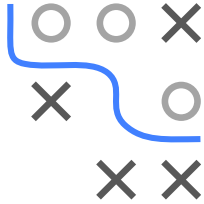


Figure 2: **Intra-class covariance.** In projection on the axes, the distributions of the two variables are the same as in the previous example. (a) The class conditional distributions have a high covariance in the direction of the line of the two class centers. There is no significant gain in separation by using two variables instead of just one. (b) The class conditional distributions have a high covariance in the direction perpendicular to the line of the two class centers. An important separation gain is obtained by using two variables instead of one.



Isabelle Guyon, André Elisseeff (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (3) p. 1157-1182.

FILTER METHODS CAN BE MISLEADING

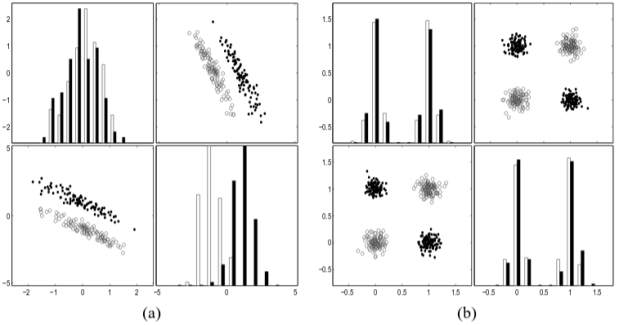
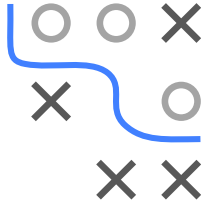
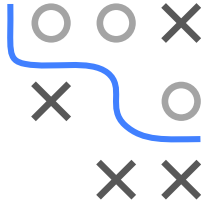


Figure 3: **A variable useless by itself can be useful together with others.** (a) One variable has completely overlapping class conditional densities. Still, using it jointly with the other variable improves class separability compared to using the other variable alone. (b) XOR-like or chessboard-like problems. The classes consist of disjoint clumps such that in projection on the axes the class conditional densities overlap perfectly. Therefore, individual variables have no separation power. Still, taken together, the variables provide good class separability .



Isabelle Guyon, André Elisseeff (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (3) p. 1157-1182.

χ^2-STATISTIC



- Test for independence between categorical x_j and cat. target y. Numeric features or targets can be discretized.
- Hypotheses:
  $H_0^j : p(x_j = m, y = k) = p(x_j = m)\, p(y = k)\, \forall m, k$
  $H_1^j : \exists\, m, k : p(x_j = m, y = k) \neq p(x_j = m)\, p(y = k)$
- Calculate the χ^2-statistic for each feature-target combination:

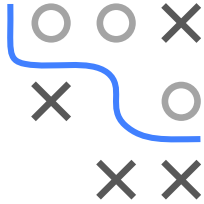
$$\chi_j^2 = \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\frac{e_{mk} - \tilde{e}_{mk}}{\tilde{e}_{mk}}\right) \quad \underset{approx.}{\overset{H_0}{\sim}} \quad \chi^2((M-1)(K-1)),$$

  where e_{mk} is the observed relative frequency of pair (m, k), $\tilde{e}_{mk} = \frac{e_{m\cdot} e_{\cdot k}}{n}$ is the expected relative frequency, and M, K are the number of values x_j and y can take.
- The greater χ_j^2, the more dependent is the feature-target combination $\rightarrow$ higher relevancy.

PEARSON & SPEARMAN CORRELATION

Pearson correlation $r(x_j, y)$:

- For numeric features and targets only.
- Most sensitive for linear or monotonic relationships.
- $r(x_j, y) = \frac{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_j^{(i)} - \bar{x}_j)}\sqrt{(\sum_{i=1}^{n} y^{(i)} - \bar{y})}}, \qquad -1 \leq r \leq 1$



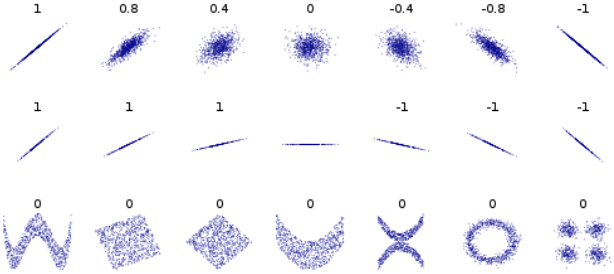
Spearman correlation $r_{SP}(x_j, y)$:

- For features and targets at least on an ordinal scale.
- Equivalent to Pearson correlation computed on the ranks.
- Assesses monotonicity of the dependency relationship.

Use absolute values $|r(x_j, y)|$ for feature ranking: higher score indicates a higher relevance.

PEARSON & SPEARMAN CORRELATION

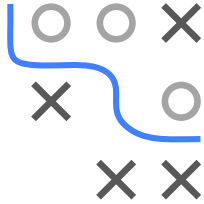
Only **linear** dependency structure, non-linear (non-monotonic) aspects are not captured:



Comparison of Pearson correlation for different dependency structures.

To assess strength of non-linear/non-monotonic dependencies, generalizations such as **distance correlation** can be used.

WELCH'S t-TEST

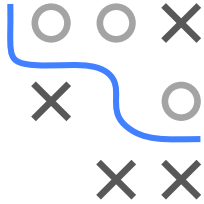


- For binary classification with numeric features.
- Statistical test for unequal means of the j-th feature.
- Let $\mathcal{Y} = \{0, 1\}$. The subscript j_0 refers to the j-th feature where $y = 0$ and j_1 where $y = 1$.
- Hypotheses:
  H_0: $\mu_{j_0} = \mu_{j_1}$ vs. H_1: $\mu_{j_0} \neq \mu_{j_1}$
- Calculate Welch's t-statistic for every feature x_j

$$t_j = \frac{\bar{x}_{j_0} - \bar{x}_{j_1}}{\sqrt{\left(\frac{S^2_{x_{j_0}}}{n_0} + \frac{S^2_{x_{j_1}}}{n_1}\right)}}$$

  where $\bar{x}_{j_y}$, $S^2_{x_{j_y}}$ and n_y are the sample mean, the population variance and the sample size for $y \in \{0, 1\}$, respectively.
- A higher t-score indicates higher relevance of the feature.

F-TEST



- For multiclass classification ($g \geq 2$) and numeric features.
- Assesses whether the expected values of a feature x_j within the classes of the target differ from each other.
- Hypotheses:
  $H_0 : \mu_{j_0} = \mu_{j_1} = \cdots = \mu_{j_g}$ vs. $H_1 : \exists\, k, l : \mu_{j_k} \neq \mu_{j_l}$
- Calculate the F-statistic for each feature-target combination:

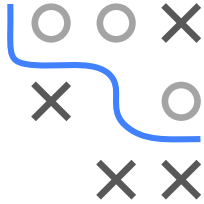
$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

$$F = \frac{\sum_{k=1}^{g} n_k (\bar{x}_{j_k} - \bar{x}_j)^2 / (g-1)}{\sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{j_k}^{(i)} - \bar{x}_{j_k})^2 / (n-g)}$$

  where $\bar{x}_{j_k}$ is the sample mean of feature x_j where $y = k$ and $\bar{x}_j$ is the overall sample mean of feature x_j.
- A higher F-score indicates higher relevance of the feature.

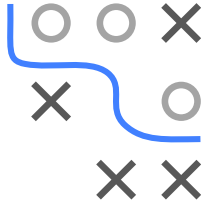
MUTUAL INFORMATION (MI)

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right]$$



- Each feature x_j is rated according to $I(x_j; y)$; this is sometimes called information gain (IG).
- MI measures the amount of "dependence" between random variables by looking how different their joint distribution is from strict independence $p(X)p(Y)$.
- MI is zero, iff $X \perp\!\!\!\perp Y$. On the other hand, if X is a deterministic function of Y or vice versa, MI becomes maximal.
- Unlike correlation, MI is defined for both numeric and categorical variables and provides a more general measure of dependence.

USING FILTER METHODS

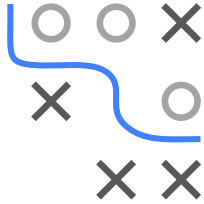


1. Calculate filter score for each feature x_j.
2. Rank features according to score values.
3. Choose $\tilde{p}$ best features.
4. Train model on $\tilde{p}$ best features.

How to choose $\tilde{p}$?

- It can be prescribed by the application.
- Eyeball estimation: read from filter plots (e.g., Scree plots).
- Use resampling.

USING FILTER METHODS



Advantages:

- Easy to calculate.
- Typically scales well with the number of features p.
- Generally interpretable.
- Model-agnostic.

Disadvantages:

- Univariate analyses may ignore multivariate dependencies.
- Redundant features will have similar weights.
- Ignores the learning algorithm.